Springboard--Data Science Career Program
Capstone Project #2 - Yelp Sentiment Analysis
Milestone Report 1
By Kevin Cole
July 2020

Problem: Positive yelp reviews are crucial for a business. Yelp not only drives traffic to business but easily helps users make decisions about which establishment to visit. These decisions are often made based on reviews and ratings found on the yelp platform.

A restaurant owner may want to tailor the customer experience to items which lead to high reviews and make business decisions to eliminate/reduce anything that can lead to negative reviews.

The goal is to use data science methods to establish a connection between the rating levels and what might drive them.

Data: The data is provided by Yelp for as an open dataset (https://www.yelp.com/dataset). This subset of Yelp data is a 10.5 GB file containing 5 different JSON encoded files. These files contain data on the business, the customer review, the user, the datetime of visit, and additional tips left by the user.

After exploration of the files, it was determined that for this project the information in the business data and the customer review data was the data of interest. 'yelp_academic_dataset_business.json' contains business info for the business in the review set. Each business is associated with some categories, information about being open/closed, hours of operation, and a unique business ID to tie the business with the appropriate reviews.

'yelp_academic_dataset_review.json' contains reviews of a large variety of businesses, their textual reviews, and the stars given by the user. This file is quite large (6.33 GB).

The business information contains a categories variable that describes all categories of business type that the business falls into. After viewing the categories in the data, the “Restaurant” category was determined to be the most valuable category tag as opposed to a more specific category like “Asian”, “Mexican”, or “Pizza”. The business information

dataset was then subsetted to include only businesses that fall into the “Restaurant” category, removing other non-food centric businesses.

The reviews dataset was then merged with the restaurant data on the unique business ID present in both datasets. Reviews for businesses not in the restaurant data were removed.

The variables of interest in the resulting dataframe were the number of stars the review gave, the review text, and a flag for the operating status of the restaurant for open/closed subsetting.

A good/neutral/bad categorical variable was created mapping 4 and 5 star ratings to good, 1 and 2 star ratings to bad, and 3 to neutral. The resulting dataset contained slightly more than 5 million Yelp reviews on restaurants. This is down from the 8 million reviews in the dataset of all reviews.

The text reviews still needed to be cleaned. The text cleaning involved the following steps.

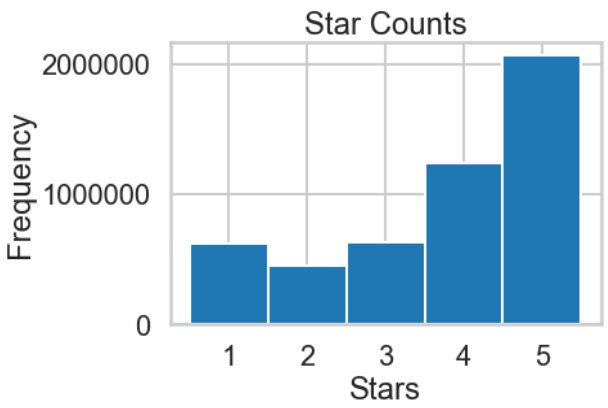
- Reviews shorter than 11 characters were deemed not valuable as the seriousness of their review is in question. These reviews were removed.
- Review text was converted to all lowercase.
- Special characters including punctuation and carriage returns were removed.
- Text language was detected using the “langdetect” python module and non-english reviews were removed. This was a computationally time intensive step.
- Stopwords defined in the NLTK python library were removed.

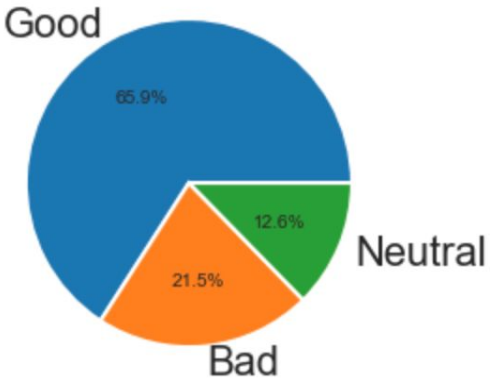
The final cleaned dataset has 5,026,116 reviews and the dataframe was saved off using the pickle format. This cleaning reduced the size of the original dataset from 10.5 GB to 1.97 GB.

	stars	text	good_bad	language	is_open
0	5	love deagans really atmosphere cozy festive sh...	Good	en	1.0
1	1	dismal lukewarm defrostedtasting texmex glop m...	Bad	en	0.0
2	4	oh happy day finally canes near casa yes other...	Good	en	1.0
3	5	definitely favorite fast food sub shop ingredi...	Good	en	1.0
4	5	really good place simple decor amazing food gr...	Good	en	1.0
...	...	...	...	...	...
5026161	5	confections cash casinos welcome las vegas fin...	Good	en	0.0
5026162	3	solid american food southern comfort flare war...	Neutral	en	1.0
5026163	5	im honestly sure never place im definitely goi...	Good	en	1.0
5026164	3	food decent say service took way long order ev...	Neutral	en	1.0
5026165	5	oh yeah service good food good serving good se...	Good	en	1.0

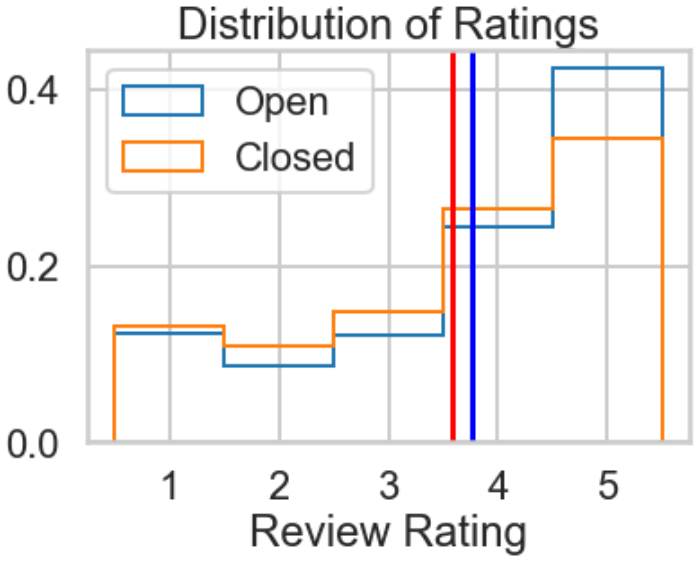
5026166 rows × 5 columns

Exploratory Data Analysis: The distribution of star counts is not normal, it is skewed towards higher ratings. This is likely due to the tendency of people to rate interactions as good unless they have a specific reason not to. The default rating tends to be 4-star or 5-star if a person has a reasonably normal interaction which produces a review set of 2/3rds of reviews classified as good reviews.





This distribution of star ratings is similar between open and closed restaurants with a noticeable drop in percentage of 5-star reviews. Restaurants receiving 5-star reviews would be less likely to close, although there are certainly other reasons for restaurants closing.



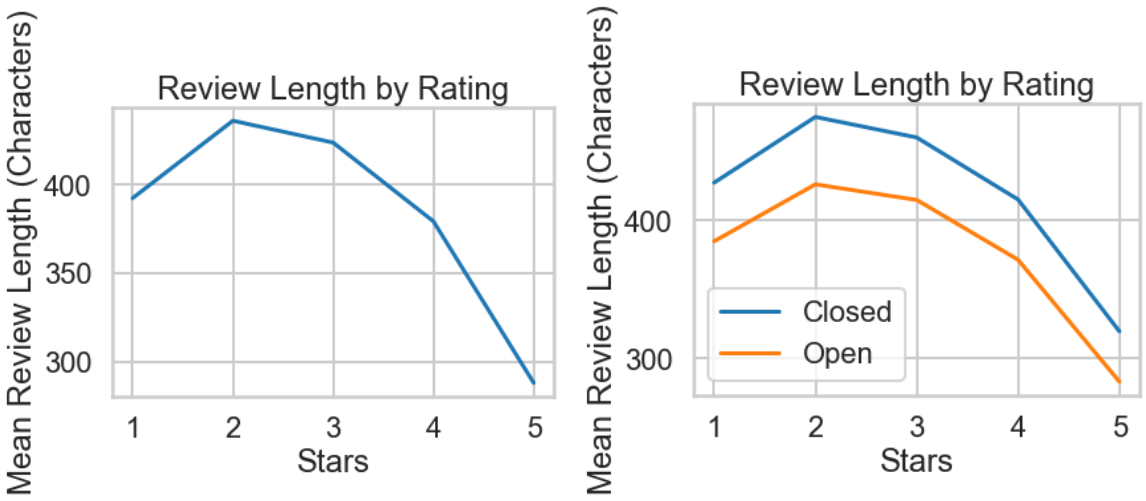
Mean rating for open restaurants: 3.76
Mean rating for closed restaurants: 3.57

Difference observed: 0.18

The two samples were used to create a bootstrapped replicate array of difference in means. In the set of replicates there was not one single occurrence in 10,000 samples of a difference in means as large as the one observed. We can conclude that this

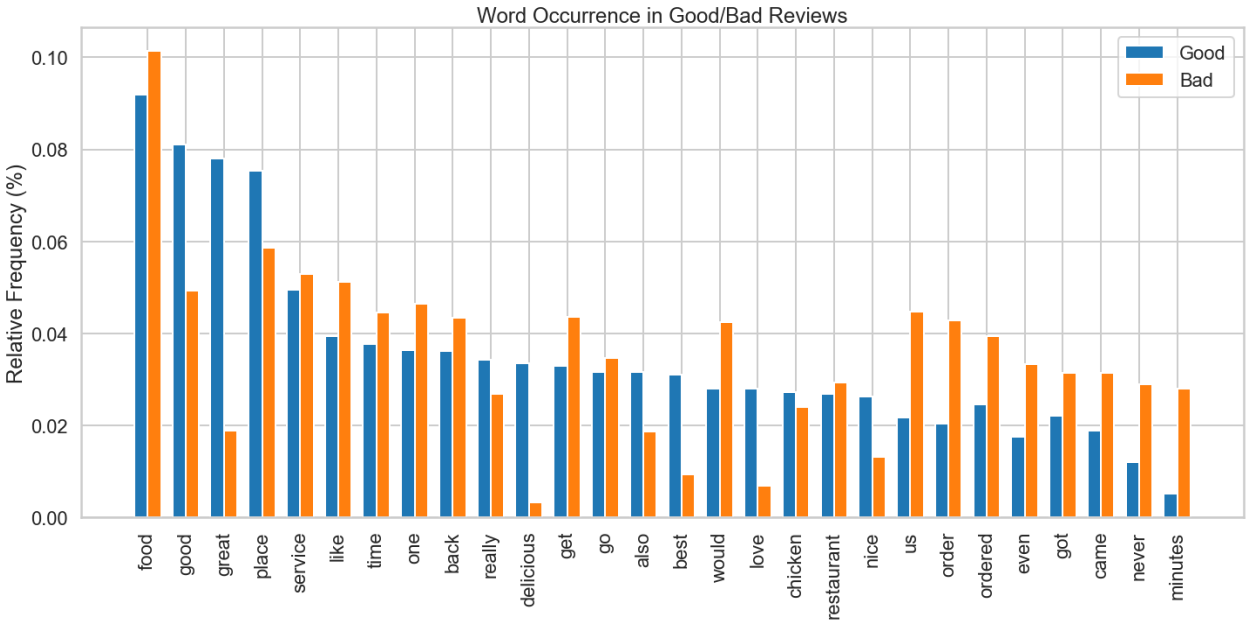
difference in the mean rating of open restaurants and closed restaurants is statistically significant.

Another key indicator to the strength of the rating is the length of the review. When people have a poor experience, they want to talk about it, and share their experience. They want to warn future patrons about what they experienced and want to elaborate, sometimes turning their review into a place to rant. For a good experience, they tend to mention a few things that they liked but do not tend to drone on and on about the experience. This tendency can be seen in the reviews of both open and closed restaurants.



The frequency of words used is a big part of natural language processing. The types of words used will play a key role in determining the strength of the review. When looking at the word frequency in both good and bad reviews we can see some predictable words and some interesting differences. Good reviews tend to be polished by strong adjectives like “good”, “great”, and “delicious”. These strong adjectives are not only prevalent in good reviews but very unlikely to occur in a bad review.

Words in negative reviews tend to be centered around service. Words like order/ordered, got/get, came, never, minutes dominate these reviews. This seems to be the largest point of complaints.



This can be seen in the word clouds created from word frequency as well.

Next steps: A supervised approach will be used, predicting which category (good/neutral/bad) each review. Classification algorithms will be explored to find what features drive each category.

In addition, an unsupervised approach will be explored with Latent Dirichlet Allocation to determine what topics are associated with good/bad reviews and to what extent those topics are correlated with their rating.